IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 7, 2018, accepted July 4, 2018, date of publication July 11, 2018, date of current version August 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2854960

Human Abnormal Behavior Impact on Speaker Verification Systems

JAROMIR TOVAREK[1], GOKHAN HAKKI ILK[2], (Member, IEEE), PAVOL PARTILA[3], AND MIROSLAV VOZNAK[3], (Senior Member, IEEE)

[1]IT4Innovations, VSB-Technical University of Ostrava, 708 00 Ostrava, Czech Republic
[2]Department of Electrical & Electronics, Faculty of Engineering, Ankara University, 06560 Ankara, Turkey
[3]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 708 00 Ostrava, Czech Republic

Corresponding author: Jaromir Tovarek (jaromir.tovarek@vsb.cz)

This work was supported by the VSB-Technical University of Ostrava, Czech Republic—Networks and Telecommunications Technologies for Smart Cities under SGS Grant SP2018/59.

ABSTRACT Human behavior plays a major role in improving human-machine communication. The performance must be affected by abnormal behavior as systems are trained using normal utterances. The abnormal behavior is often associated with a change in the human emotional state. Different emotional states cause physiological changes in the human body that affect the vocal tract. Fear, anger, or even happiness we recognize as a deviation from a normal behavior. The whole spectrum of human-machine application is susceptible to behavioral changes. Abnormal behavior is a major factor, especially for security applications such as verification systems. Face, fingerprint, iris, or speaker verification is a group of the most common approaches to biometric authentication today. This paper discusses human normal and abnormal behavior and its impact on the accuracy and effectiveness of automatic speaker verification (ASV). The support vector machines classifier inputs are Mel-frequency cepstral coefficients and their dynamic changes. For this purpose, the Berlin Database of Emotional Speech was used. Research has shown that abnormal behavior has a major impact on the accuracy of verification, where the equal error rate increase to 37 %. This paper also describes a new design and application of the ASV system that is much more immune to the rejection of a target user with abnormal behavior.

INDEX TERMS Abnormal behavior, emotion, voice, verification, SVM.

I. INTRODUCTION

Exchange of information is inherently linked to the mutual identification of communicating participants. An ordinary conversation between two people always begins with the identification of both sides and continues by mutual trust. The process of simplifying human-machine communication logically seeks user-friendly identification and verification methods. Therefore, speaker verification presents a significant challenge in advancing the current technological trend. This fact is visible in speech technologies used by the public and commercial spheres. Compared to other methods of biometric verification (e.g. fingerprint, iris, facial), human speech contains one significant advantage. In addition to the content, speech includes information about speakers (age, gender, emotion, but mainly identity).

Speech verification also has its weaknesses. Speech is pronounced differently in different situations. The main reasons are the influence of emotions and body response on the vocal tract. Allen *et al.* [1] present the results of an experiment demonstrating the impact of external stress stimuli on the cardiac and respiratory activity of the human body. It is also well known that heartbeat and breathing are influenced by psychological stimuli, which leads to active emotions [2], [3]. Due to the impact of emotional states on the human body, Cowie and Cornelius [4] distributes emotions actively and passively (see Fig. 1). Based on the above mentioned knowledge, we recognize passive emotions (neutral, sadness, boredom) as the normal speaker behavior and active emotions (anger, fear, happiness) as the abnormal behavior of a speaker that may affect ASV accuracy. Each of these emotions listed below can be encountered by an ASV system and considered abnormal in this study because traditional ASV systems disregard them. The main question of interest in this study is whether these emotional states, not considered by a traditional ASV, affects the performance of the system or not. These intra-speaker variabilities resulted from emotional state of the

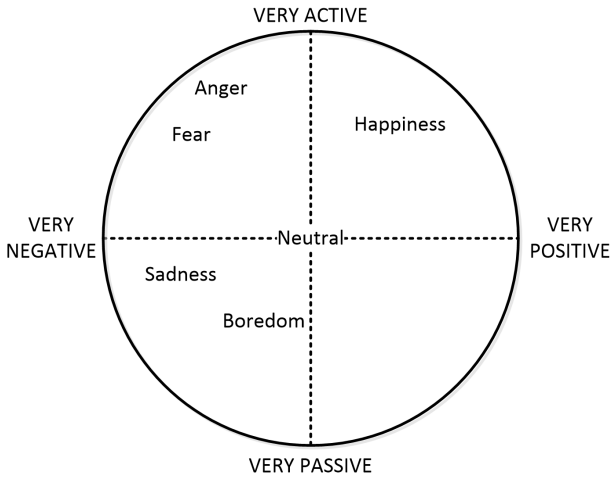

FIGURE 1. Circle of emotions. Emotional states are located in space reference to the neutral state and their character. Using this graph, active and passive emotional states can be divided [4].

speaker are considered abnormal and they are held distinct from effects like vocal aging, disguised voices and health related issues etc. because the target users were considered cooperative but under certain emotions beyond their control.

- Anger (the user may get angry simply because he/she did not get access granted on his/her previous trial although he/she is not an imposter)
- Happy (the user may be feeling happy just because of everyday life, daily life pleasures)
- Fear (the user may be under psychic tension about the possibility of not getting access granted)

Related Work

The impact of abnormal behavior and emotional speech on speaker recognition and verification has been the subject of several studies [4]–[8]. The cited works use various techniques to improve the accuracy or problem definition. MFCC and its dynamical changes were used for obtaining promising results [9]. There are many classification methods for speaker verification [10]. SVM has been selected for this task as it is a binary classifier that can achieve promising results on a small amount of data as this is case of our study [11].

As mentioned above, the pronunciation method directly affects the accuracy and effectiveness of the ASV systems. Excluding physiological changes of the vocal tract caused by the current disease (e.g., influenza, angina, and others), an emotional state change represents abnormal behavior of the verifying participant [12]–[14]. Many databases are available and used for training and testing of ASV systems (M2VTS [15], XM2VTS [16], RSR2015 [17], SAS [18]). Detailed list of available databases is presented in work of Larcher *et al.* [19]. Most of them were designed to eliminate the issue of spoofing and other verification weaknesses which are beyond the scope of this paper [20], [21].

This article addresses the influence of abnormal behavior on the accuracy of the speaker verification. Therefore, an emotionally colorful speech recordings had to be used. For this purpose, Berlin Database of Emotional Speech [22] has been selected, which represents one of the reference publicly available databases for recognizing the emotional state of human speech in recent years [23].

The following chapters describe methods used to design the ASV system in our study. The first part is an experiment that examines degradation of verification accuracy and effectiveness for abnormal behavior of the speaker.

II. METHODS

The process of voice authentication consists of following steps: signal preprocessing, feature extraction, classification, evaluation and decision. These steps are shown in Fig. 2.

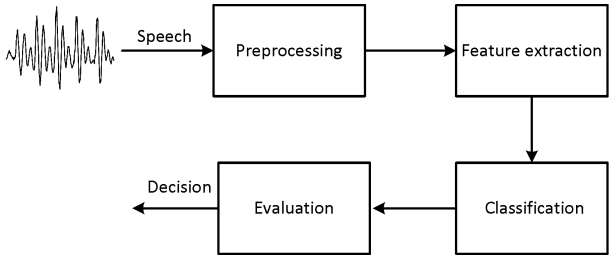

FIGURE 2. Process of voice authentication.

A. SIGNAL PREPROCESSING

Signal preprocessing is the first step of voice authentication. Preprocessing performs an adjustment of the speech signal into a useful form. Usually, preprocessing consists of five operations: removal of DC offset, pre-emphasis, segmentation, smoothing function and removal of silence. First four operations are described in [24]. As we want to model the user's speech and not the type of silence. It is common to remove the silent segments. Voice Activity Detection (VAD) based on low energy segments is applied for this purpose [25].

B. FEATURE EXTRACTION

The most important step of voice authentication is the choice of significant parameters/features. These parameters should meet some requirements. First, the parameters should occur in speech commonly thus making the measurement easy. Second, they should be robust. Parameters should not change their characteristics in time or under varying health conditions. Third, they should be secure, which means that it is not easy to mimic these parameters [26], [27]. The extraction of MFCCs and their derivations (delta and delta-delta coefficients) is a very common choice in the field of speaker recognition [28], [29]. We used thirteen MFCC coefficients and their derivatives without the first coefficient (c_0). This coefficient carries information only about signal energy.

1) MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Human hearing is non-linear. This feature is compensated by triangular filters with nonlinear frequency distribution defined by (1). Mel filter bank is used in calculating the MFCC, which are defined by (2). Dynamic coefficients

denoted as delta and delta-delta (acceleration coefficients) represent the dynamic time variations (derivation) of the MFCC [26], [29]–[31].

$$f_m = 2595 \log\left(1 + \frac{f}{700}\right), \tag{1}$$

where f is the frequency in hertz scale and f_m is frequency in mel scale.

$$c_m(j) = \sum_{i=1}^{M^*} \log y_m(i) \cos\left(\frac{\pi j}{M^*}(i - 0,5)\right), \quad for\ j = 0, 1, \ldots, M, \tag{2}$$

where $y_m(i)$ is the filter response, M^* represents a number of bands in the filter bank, and M is the number of cepstral coefficients.

C. CLASSIFICATION

The next step after feature extraction is classification. We used SVM approach. SVM offers a progressive method in the field of machine learning. This approach is primarily intended for binary classification [31]. A simple Min-max normalization is applied before classification [33].

1) SUPPORT VECTOR MACHINES

The principle of classification is to find the hyperplane that divides the training data in the feature space as shown in Fig. 3. The optimal hyperplane is such that the training data points lie in the opposite half-space and the value of the distance between half-spaces is the largest. In other words, the goal is to maximize space among half-spaces (maximum margin). Support vectors are described by training data points that represent a decision-making role [28], [31], [32].

The training process is based on minimalization of weigths vector size defined by (3) with condition (4).

$$|\vec{w}_{SVM}| = \frac{1}{2}\vec{w}_{SVM}^T \vec{w}_{SVM}, \tag{3}$$

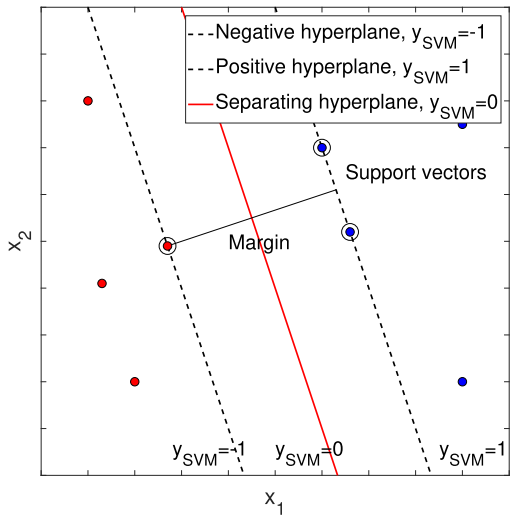

FIGURE 3. Principle of SVM.

where $\vec{w}_{SVM}$ is a vector of SVM classifier weights and T denotes the transpose.

$$t_n(\vec{w}_{SVM}^T \vec{X}_i + b) \geq 1, \quad for\ i \in \{1, 2, \ldots, K\}, \tag{4}$$

where $t_n \in \{-1, 1\}$ are classes for training data, b is a bias, $\vec{X}_i$ is a vector of training data and K is the total number of training data.

Vector of weights can be expressed as dot product of training data described by (5). The dot product is represented by kernel functions for high dimensional space [31]. The features are not separable by a linear function and therefore we used nonlinear (quadratic, cubic and quartic) and Gaussian kernels in the classifier. Best and satisfactory results are obtained by the cubic kernel function.

$$\vec{w}_{SVM} = \sum_{i \in K_P} t_i \beta_i \vec{X}_i, \quad where\ \beta_i > 0, \tag{5}$$

where K_P is the total number of support vectors, β_i is a weight of support vector and t_i is a class of support vector.

The classification is performed by calculating the dot product between the support vectors and the test data vectors (defined below by (6)).

$$y_{SVM} = \sum_{i \in K_P} t_i \beta_i \vec{X}_i \vec{X}_{test} + b, \tag{6}$$

where y_{SVM} is a score of SVM classifier and $\vec{X}_{test}$ is a vector of testing data.

The use of SVM on acoustic features, plus their derivatives extracted from 20 ms segments (frames) in this study also differs from conventional GMM based models. This approach is used in order to achieve promising results on a small amount of data.

D. EVALUATION

SVM generates output for each speech segment. The output of this classifier is posterior probabilities for both classes. Sum of these two probabilities is equal to 1. Posterior probability represents the probability that the segment belongs to target user or imposter.

After classification, each segment is evaluated separately, but the aim is the evaluation of the whole record. Therefore the approach of the information fusion with majority vote rule is applied [34].

Majority vote rule is obtained from the sum rule [35]. The first step of this method is an approximation of posterior probabilities. Approximation is defined by (7). After this, we can simply sum the votes for both classes on the right side of (8) and we can compare the max value with a threshold. If the max value is lower than a threshold, the winning speaker has marked an imposter.

$$\Delta_{ki} = \begin{cases} 1 & \text{if } P(\omega_k|\vec{x}_i) = \max_{j=1}^{m} P(\omega_j|\vec{x}_i) \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $P(\omega_k|\vec{x}_i)$ is the highest probability from all output classes for feature vector $\vec{x}_i$, $P(\omega_j|\vec{x}_i)$ is the posterior probability for output class ω_j and m is the total number of output classes.

$$\sum_{i=1}^{R} \Delta_{ji} = \max_{k=1}^{m} \sum_{i=1}^{R} \Delta_{ki}, \tag{8}$$

where R is the total number of segments where the posterior probabilities was maximum for output class k throughout the recordings.

E. DECISION

The last step of the authentication process is the decision about authenticating. The system has to make a decision whether the user is target user or an imposter. The decision is based on a comparison of a max value of score and threshold. If the max value is higher than a threshold, the speaker is marked as target user otherwise as an imposter.

Measurement of voice authentication performance allows comparison of different systems. We used false acceptance rate (FAR), false rejection rate (FRR) , equal error rate (EER), receiver operating characteristic (ROC) and detection error tradeoff (DET) curves for measurement of performance. In the field of verification systems, the false rejection is often called ''miss'' and false acceptance is also called ''false alarm''.

Achieved results are also presented by confusion matrix. The rows correspond to the predicted class (Output Class) and the columns correspond to the actual (Expected Class). The diagonal cells correspond to percentage of attempts that are correctly classified. The off-diagonal cells correspond to incorrectly classified observations. The column on the far right side of the table shows the percentages of all the examples predicted to belong to each class that are correctly classified. These metrics are often called the precision (or positive predictive value). The row at the bottom of the table shows the percentages of all the examples belonging to each class that are correctly classified. These metrics are often called the recall (or true positive rate). The cell in the bottom right of the table shows the overall accuracy.

The FAR is the measure of the likelihood that the voice authentication system will incorrectly accept an access attempt by the imposter. FAR is computed by (9). The FRR is the measure of the likelihood that the voice authentication system will incorrectly reject an access attempt by a target user. FRR is computed by (10). EER indicates that the proportion of FAR is equal to the proportion of FRR. The threshold value for EER is called equal error threshold (EET). ROC shows the relationship between true positive rate (TPR) and false positive rate (FPR) at various threshold settings. DET curve is a graphic representation of error rates (FAR vs FRR) for binary classification systems [28], [36].

$$FAR = \frac{N_{FA}}{N_{IVA}}, \tag{9}$$

where N_{FA} is the number of incorrect acceptance (false alarm) and N_{IVA} is the number of all imposter attempts.

$$FRR = \frac{N_{FR}}{N_{EVA}}, \tag{10}$$

where N_{FR} is the number of incorrect rejection (miss) and N_{EVA} is the number of all authorized attempts.

III. EXPERIMENT AND RESULTS

The experiment was focused on verifying the hypothesis that voice authentication accuracy is affected by speech signal obtained from abnormal behavior. Results show the comparison of ASV system accuracy for speech of normal state and abnormal state (behavior). The Berlin database of Emotional Speech is used [22]. The recordings are divided into groups of normal behavior and abnormal behavior. The emotions contained in the database are defined by the circle of emotion [4]. This 2D model divides emotion states by psychological impact to active and passive. Normal behavior consists of neutral, boredom and sadness and abnormal behavior consists of anger, fear and happiness [4].

The experiment is divided into two parts. The first part (Sec. III-A) evaluates the effect of users' speech in an abnormal state to ASV systems, resulting in a degradation of the system accuracy. The second part of the experiment (Sec. III-B) describes the proposal for system improvement.

A. VERIFYING THE IMPACT OF ABNORMAL BEHAVIOUR

The text-independent ASV system was trained for verification of 10 target users (5 women, 5 men). SVM classifier is trained traditionally by normal state recordings. New user comes to registration process of verification without abnormal symptoms (stress, psychological pressure, mood). Therefore, the system is trained by normal state recordings. Speech parameters of one target user represent class 1 (class 1 – verified user). Background model represents imposter, where rest of nine users are used for training (class 2 – for each unverified user, an imposter). One SVM model represents one target user (10 models for 10 target users). The system was evaluated in two testing phases. At first, normal behavior testing phase was evaluated by classification (verification) of 200 recordings. The system was tested by 10 target users (100 recordings) and 10 imposters (100 recordings), both in normal behavior. In the second phase, system accuracy is evaluated by 200 abnormal behavior recordings. The ratio of target users and imposters is the same as in the first phase but none of the recordings that have been used for training is used during the test. We have trained SVM on the frame-level acoustic features and then forming a majority vote style classification rule to combine the frame-level decisions into a sequence-level decision. The objective of the experiment is a comparison of EER and system accuracy for users verified with normal and abnormal behavior. For the first phase system achieved 4 % of EER with 46.0 % EET. In the second phase, the system was tested by abnormal behavior recordings. The system achieved 37 % of EER with 26.9 % EET.

These results confirm the assumption of reduced accuracy and system effectiveness for the user in the abnormal state.

B. IMPROVEMENT OF THE SYSTEM

As was mentioned in the introduction, a human often gets into a situation where he is influenced by different stimuli. This view of the issue was the reason for including a new class for classification.

The new system design includes a classifier trained by three classes. The aim is to extend the existing verification system to the ability to recognize user under abnormal behavior. SVM classifier was trained using 100 recordings per class. The training rate is the same as in the first part of the experiment (Sec. III-A). The same ratio was used for system testing. Classes are now defined as:

- class 1 - target user in normal behavior,
- class 2 - target user in abnormal behavior,
- class 3 - imposter.

Table 1 represents the confusion matrix of the newly designed system. The confusion matrix clearly shows that 96% of normal target users are correctly predicted with only 4% confusion (miss-classification) with abnormal target users. There is no confusion with imposters for prediction of normal target users. Likewise for prediction of abnormal target users the classification rate is 95% with 4% confusion with normal target users and 1% with imposters. Last one imposter class is only confused with normal target users (4%) indicating that normal and abnormal behavior (classes) are distinct.

TABLE 1. The confusion matrix of proposed verification system trained with three classes. Percentage of attempts classified for each class are shown in each cell (Sec.:II-E describes in detail). Results are based on the majority voting method.

Predicted classes	*Target users - Normal*	96%	4%	4%	92.3%
	Target users - Abnormal	4%	95%	0%	96.0%
	Imposters	0%	1%	96%	99.0%
		96.0%	95.0%	96.0%	**95.7%**
		Target users - Normal	*Target users - Abnormal*	*Imposters*	
		Actual classes			

Information about the behavior of the user during verification offers two options. The first option points to the additional verification process. Target users with abnormal behavior are not strictly rejected - the system notifies the user and offers an additional verification procedure such as next attempt for speech verification or another verification method.

The second option points to information about abnormal behavior and its application directly in the first verification attempt. From the view of granting or denying user access to the verification process, information about its behavior is not important. The only important and final decision is granting or denying access. For this reason, we can sum posterior probabilities of the first class ω_1 and second class ω_2. The new rule we propose in this paper regardless of normal or abnormal target user is:

$$\text{Dec.} = \begin{cases} 1 & \text{if } P(\omega_1|X_i) + P(\omega_2|X_i) > \text{thrs} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

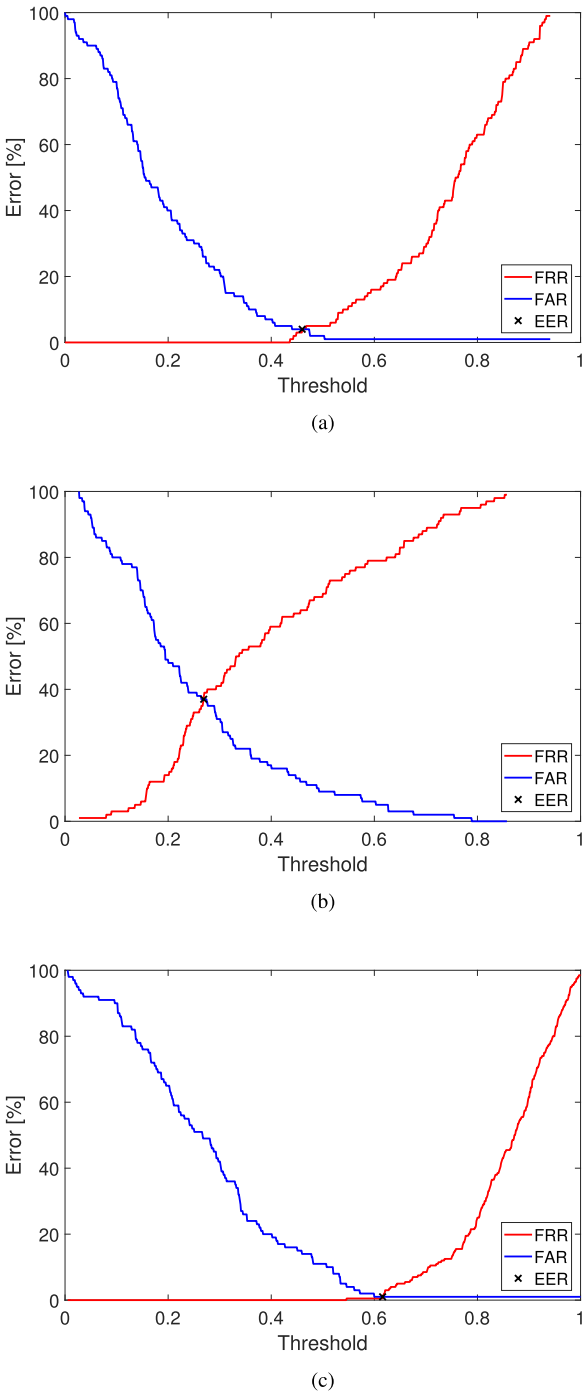

FIGURE 4. ASV system errors for different threshold. Graphs 4(a), 4(b) and 4(c) describe relation between errors (FAR and FRR) and threshold for different system approaches. The curve penetration defines the EER at the corresponding EET value. (a) Classical approach - normal behavior. (b) Classical approach - abnormal behavior. (c) New approach.

The result is the posterior probability of granting access. In this case, the new proposed system has reached 1% EER for 61.6% EET. This system achieved 99% of accuracy. Table 2 shows a comparison of the classical approaches with the new proposed approach. The lowest EER value prove the benefit of the proposed system. The dependence of FAR and FRR errors on the threshold for all studied approaches are shown in the graphs in Fig. 4.

TABLE 2. Systems comparison with EER and EET values. The new approach reaches best (the lowest) EER value.

Systems	EER [%]	EET [%]
Classical approach (normal behaviour)	4	46.0
Classical approach (abnormal behaviour)	37	26.9
New approach	1	61.6

Another visual comparison of the results is shown using ROC and DET curves. Fig. 5 and 6 show the comparison of presented systems. From both curves, it is obvious that the new approach has achieved the best results (the ROC curve approaches the ideal course - Area under the curve (AUC) is close to 1). Conversely, the course of the DET curve is tilted to zero FAR and FRR values.

Data collection - re-training

The presented new proposal has one key problem. How to train a system to detect an abnormal state when a user has registered with a speech in a normal state? In other words, registering a user into the verification system does not provide data for the training of all three classes.

Solving this problem may be easier than it seems. Verification systems are in most cases deployed as part of a multi-phase verification. The verification process during user denial often offers additional access opportunities, such as PIN, password, or other biometric methods. Finally, the rejected user's speech can be considered as abnormal user behavior if access was granted in another verification phase. This speech can be used to re-configure and re-train the system into the design of the proposed approach.

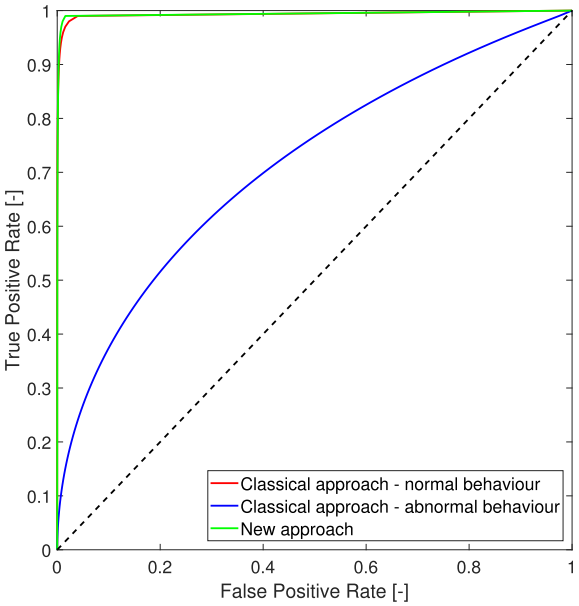

FIGURE 5. ROC curves - comparison of classical approach (Sec. III-A) and new proposed system (Sec. III-B) tested on users with normal and abnormal behavior.

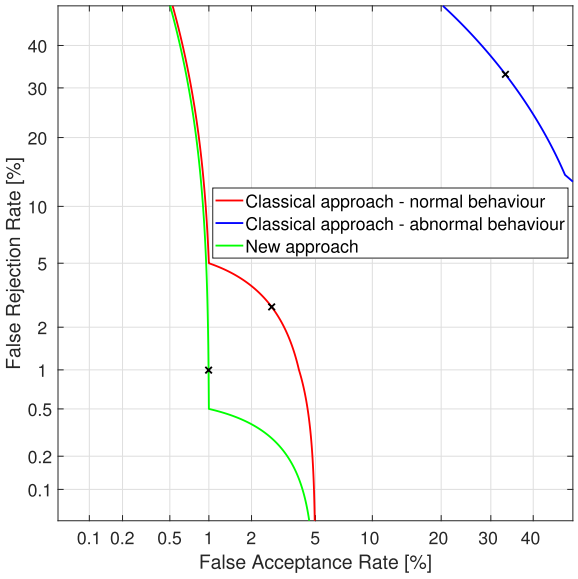

FIGURE 6. DET curves - comparison of presented approaches (same in Fig. 5). The points on the curves represents EER for each approach.

IV. CONCLUSION AND DISCUSSION

The research was aimed at the impact of abnormal behavior on the ASV systems. Active emotions, which are influenced by psychological and physiological changes of the vocal tract are consider as abnormal. The first part of the experiment demonstrated the undesirable influence of abnormal behavior on the accuracy and effectiveness of the verification system trained with users under normal behavior. Ten target users were used for the verification process. Ten SVM models were trained by speech recordings (MFCC and their dynamic changes) of users in a normal state (behavior). The accuracy of verification system was evaluated with normal versus abnormal behavior users. The system EER increases from 4 % to 37 % for speech represented by users with abnormal behavior. These results confirm the adverse impact of human abnormal behavior on voice authentication accuracy. The fact that abnormal behavior increases EER value means that the system declines target users under abnormal behavior. Two preconditions and reasons derive from experimental results. At first, the threat of system security because a target user under abnormal behavior may be forced by a third person to grant access, or second, the system rejects the target user only for being embarrassed by normal life stimuli. The second part of the research brings a new design of the system and the way of its application. The new proposal involves recognizing three states (classes), namely: i. target user (normal behavior),

ii. target user (abnormal behavior), and iii. imposter. The final decision to grant or deny access depends on the new rule defined by (11). The application of the presented design and solving the problem of data collection (abnormal behavior speech data) are so presented.

The advantage of the new design is the improvement of accuracy for verification of users with abnormal behavior. The fact that new proposal can be applied to existing systems is also a major contribution of our research.

REFERENCES

[1] M. T. Allen, A. Sherwood, and P. A. Obrist, "Interactions of respiratory and cardiovascular adjustments to behavioral stressors," *Psychophysiology*, vol. 23, no. 5, pp. 532–541, 1986, doi: 10.1111/j.1469-8986.1986.tb00669.x.

[2] D. Carroll, J. R. Turner, and J. C. Hellawell, "Heart rate and oxygen consumption during active psychological challenge: The effects of level of difficulty," *Psychophysiology*, vol. 23, no. 2, pp. 174–181, 1986, doi: 10.1111/j.1469-8986.1986.tb00613.x.

[3] S. Johar, *Emotion, Affect and Personality in Speech: The Bias of Language and Paralanguage* (SpringerBriefs in Electrical and Computer Engineering), vol. 52. Cham, Switzerland: Springer, 2016, pp. 9–15, doi: 10.1007/978-3-319-28047-9.

[4] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, nos. 1–2, pp. 5–32, 2003, doi: 10.1016/S0167-6393(02)00071-7.

[5] W. Wu, T. F. Zheng, M.-X. Xu, and H.-J. Bao, "Study on speaker verification on emotional speech," in *Proc. INTERSPEECH*, Pittsburgh, PA, USA, 2006, pp. 2102–2105.

[6] P. Staroniewicz, "Influence of speakers' emotional states on voice recognition scores," in *Proc. Anal. Verbal Nonverbal Commun. Enactment, Process. Issues*, Budapest, Hungary, 2010, pp. 223–228.

[7] T. Wu, Y. Yang, and Z. Wu, "Improving speaker recognition by training on emotion-added models," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Beijing, China, 2005, pp. 382–389.

[8] L. Chen and Y. Yang, "Emotional speaker recognition based on model space migration through translated learning," in *Proc. Chin. Conf. Biometric Recognit.*, Jinan, China, 2013, pp. 394–401.

[9] A. Mansour and Z. Lachiri, "SVM based emotional speaker recognition using MFCC-SDC features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 538–544, 2017, doi: 10.14569/IJACSA.2017.080471.

[10] S. S. Admuthe and S. Ghugardare, "Survey paper on automatic speaker recognition systems," *Int. J. Eng. Comput. Sci.*, vol. 4, no. 3, pp. 10895–10898, 2015.

[11] Z. Cirovic and N. Cirovic, "A robust SVM/GMM classifier for speaker verification," in *Proc. Int. Conf. Speech Comput.*, Novi Sad, Serbia, 2014, pp. 74–80.

[12] C. E. Izard, E. A. Youngstrom, S. E. Fine, A. J. Mostow, and C. J. Trentacosta, "Emotions and developmental psychopathology," in *Developmental Psychopathology: Theory and Method*, D. Cicchetti and D. J. Cohen, Eds. Hoboken, NJ, USA: Wiley, 2015, pp. 244–292.

[13] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015, doi: 10.1007/s10462-012-9368-5.

[14] M. S. Hossain and G. Muhammad, "An emotion recognition system for mobile applications," *IEEE Access*, vol. 5, pp. 2281–2287, 2017, doi: 10.1109/ACCESS.2017.2672829.

[15] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," in *Audio- and Video-based Biometric Person Authentication*, J. Bigun, G. Chollet, and G. Borgefors, Eds. Berlin, Germany: Springer, 1997, pp. 403–409.

[16] K. Messer, J. Matas, J. Kittler, and K. Johsson, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, Washington, DC, USA, 1999, pp. 72–77.

[17] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 1580–1583.

[18] Z. Wu *et al.*, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4440–4444, doi: 10.1109/ICASSP.2015.7178810.

[19] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Commun.*, vol. 60, pp. 56–77, May 2014, doi: 10.1016/j.specom.2014.03.001.

[20] Z. Wu *et al.*, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017, doi: 10.1109/JSTSP.2017.2671435.

[21] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of Biometric Anti-Spoofing*, S. Marcel, M. S. Nixon, and S. Z. Li, Eds. London, U.K.: Springer-Verlag, 2014, pp. 125–146.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520.

[23] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.

[24] J. Tovarek, P. Partila, M. Voznak, M. Mikulec, and M. Mehic, "Detection of cardiac activity changes from human speech," *Proc. SPIE*, vol. 9496, p. 94960V, May 2015.

[25] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *Proc. 6th Int. Conf. Signal Process.*, 2002, pp. 464–467, doi: 10.1109/ICOSP.2002.1181092.

[26] S. Debnath, B. Soni, U. Baruah, and D. K. Sah, "Text-dependent speaker verification system: A review," in *Proc. ISCO*, Coimbatore, India, 2015, pp. 1–7.

[27] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010, doi: 10.1016/j.specom.2009.08.009.

[28] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015, doi: 10.1109/MSP.2015.2462851.

[29] Z. Ma, H. Yu, Z.-H. Tan, and J. Guo, "Text-independent speaker identification using the histogram transform model," *IEEE Access*, vol. 4, pp. 9733–9739, 2017, doi: 10.1109/ACCESS.2016.2646458.

[30] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande, "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network," in *Proc. ICAPR*, Kolkata, India, 2015, pp. 1–6.

[31] S.-H. Chen and Y.-R. Luo, "Speaker verification using MFCC and support vector machine," in *Proc. IMECS*, Hong Kong, 2009, pp. 1–4.

[32] Z. Sun, Z. Guo, C. Liu, X. Wang, J. Liu, and S. Liu, "Fast extended one-versus-rest multi-label support vector machine using approximate extreme points," *IEEE Access*, vol. 5, pp. 8526–8535, 2017, doi: 10.1109/ACCESS.2017.2699662.

[33] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005, doi: 10.1016/j.patcog.2005.01.012.

[34] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998, doi: 10.1109/34.667881.

[35] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Anal. Appl.*, vol. 1, no. 1, pp. 18–27, 1998, doi: 10.1007/BF01238023.

[36] M. El-Abed and C. Charrier, "Evaluation of biometric systems," in *New Trends and Developments in Biometrics*, J. Yang and S. J. Xie, Eds. Rijeka, Croatia: InTech, 2012, pp. 149–169.

JAROMIR TOVAREK received the Ph.D. degree in telecommunications in 2018. His thesis was focused on the multimodal biometric authentication systems. He is currently a Researcher with the IT4Innovations, VSB-Technical University of Ostrava, Czech Republic. His research is focused on signal processing, face recognition, speaker recognition, and speech recognition. He is a member with the INTERSPEECH Conference Committee.

GOKHAN HAKKI ILK (M'00) was born in Ankara, Turkey, in 1971. He received the B.Sc. degree from Ankara University, Ankara, in 1993, the M.Sc. degree in instrument design and applications from the University of Manchester Institute of Science and Technology, in 1994, and the Ph.D. degree from the University of Manchester, Manchester, U.K., in 1997.

He is currently a Professor with the Electrical and Electronics Engineering Department, Ankara University. He has 74 publications, according to Google scholar and 297 citations. His publications are in bioinformatics, optical communications, digital speech, and image signal processing. His current research interests are digital signal processing, namely speech, image, and video processing.

Dr. Ilk has a book in Turkish on *Applied Signal Processing*. He is the Founder of the Ankara University Speech Processing Group and shared Turkcell's (biggest GSM operator in Turkey) Best Academic Study Award in 2007.

PAVOL PARTILA received the Ph.D. degree for a thesis which focused on the issues of speech emotion recognition systems in 2017. He is currently a Post-Doctoral Researcher with the Department of Telecommunications, VSB-Technical University of Ostrava, Czech Republic. His research is focused on speech processing, speech quality, voice over Internet Protocol, and Internet of Things. He is a member with the INTERSPEECH Conference Committee.

MIROSLAV VOZNAK (M'10–SM'16) received the Ph.D. degree in telecommunications and the Habilitation degree from the Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, in 2002 and 2009, respectively. He was appointed as a Full Professor in electronics and communication technologies in 2017. His research interests focus generally on information and communications technology, particularly on quality of service and experience, network security, wireless networks, and in the last couple years also on big data analytics in mobile cellular networks. He is a member in numerous IEEE conference committees and he has served as a member of the editorial board for several journals, such as the *Journal of Communications* or a Guest-Editor of the *Wireless Personal Communications*.